W101

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2001-014326

(43)Date of publication of application : 19.01.2001

(51)Int.Cl. G06F 17/30
G06F 17/27
G06F 17/21

(21)Application number : 11-183349

(22)Date of filing : 29.06.1999

(71)Applicant : HITACHI LTD

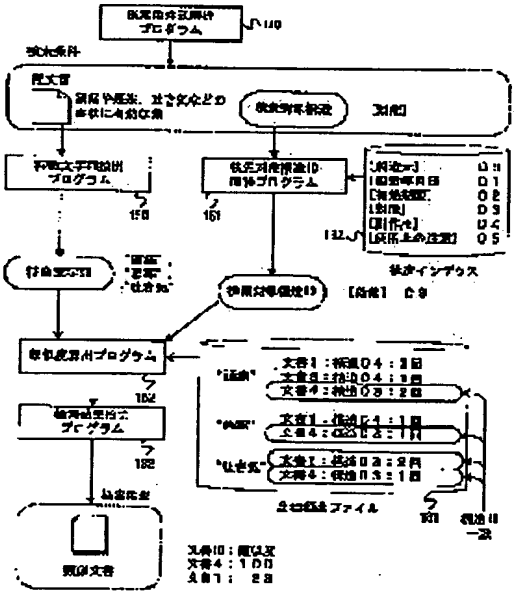
(72)Inventor : MATSUBAYASHI TADATAKA
TADA KATSUMI
SUGAYA NATSUKO
INABA YASUHIKO
YAMAGUCHI AKIHIKO
GOCHI YOSUKE

(54) DEVICE AND METHOD FOR RETRIEVING SIMILAR DOCUMENT BY STRUCTURE SPECIFICATION

(57)Abstract:

PROBLEM TO BE SOLVED: To add the specification of an object structure to be retrieved to retrieval conditions and to improve the retrieval precision when a document which is similar to a seed document (document specified as a retrieval condition) is retrieved.

SOLUTION: A retrieval condition expression analyzing program 130 receives the specification of a seed document and the input of an object structure to be retrieved as retrieval conditions. A featured character string extracting program 150 extracts a featured character string from the text of the specified seed document. A retrieval object structure ID acquiring program 151 converts the specified structure into its ID. A similarity calculating program 152 performs retrieval from an appearance frequency file 181 to acquire the appearance frequency of a document whose structure ID matches the featured character string and calculates the similarity of the similar document based upon the seed document. A retrieval result output program 132 displays the identifier and similarity of the similar document as the retrieval result.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－14326
（P2001－14326A）
(43)公開日　平成13年１月19日(2001.1.19)

(51)Int.Cl.'	識別記号	ＦＩ		テーマコード'（参考）
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３５０Ｃ	５Ｂ００９
17/27		15/20	５５０Ｅ	５Ｂ０７５
17/21			５９０Ｅ	
		15/40	３４０	
			３７０Ａ	

審査請求　未請求　請求項の数22　ＯＬ　（全 16 頁）

(21)出願番号　特願平11－183349

(22)出願日　平成11年６月29日(1999.6.29)

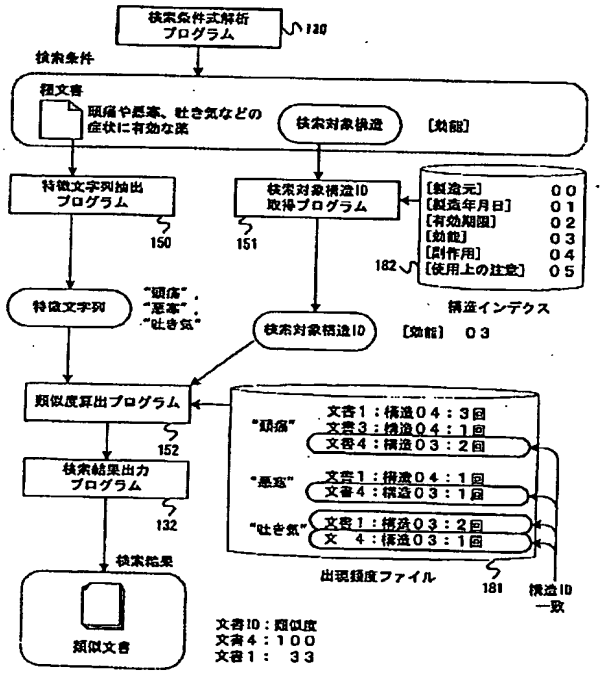
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　松林　忠孝
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(72)発明者　多田　勝己
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(74)代理人　100061893
弁理士　高橋　明夫　（外１名）

最終頁に続く

(54)【発明の名称】　構造指定による類似文書の検索装置及び検索方法

(57)【要約】
【課題】　種文書に類似する文書を検索するに際して、検索条件に検索対象構造の指定を付加し、もって検索精度の向上を図る。
【解決手段】　検索条件式解析プログラム１３０は、検索条件として種文書の指定と検索対象構造の入力を受ける。特徴文字列抽出プログラム１５０は、指定された種文書のテキストから特徴文字列を抽出する。検索対象構造ＩＤ取得プログラム１５１は、指定された構造をそのＩＤに変換する。類似度算出プログラム１５２は、出現頻度ファイル１８１を検索して特徴文字列と構造ＩＤが一致する文書の出現頻度を取得し、類似文書の種文書を基とする類似度を計算する。検索結果出力プログラム１３２は、検索結果の類似文書の識別子とその類似度を表示する。

図2



【特許請求の範囲】

【請求項１】計算機を利用して検索条件として指定された文書あるいは文章（以下まとめて種文書と呼ぶ）に類似する構造化文書を検索する方法であって、類似度計算の検索条件として種文書と該構造化文書に属する少なくとも１つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項２】計算機を利用して種文書に類似する構造化文書を検索する方法であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストから特徴となる文字列を抽出するステップと、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として、該特徴文字列に基づく該種文書との類似度を算出するステップと、算出された類似度の高い順に従って表示の優先度を決定するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項３】計算機を利用して構造化された種文書に類似する文書を検索する方法であって、種文書と該種文書に属する少なくとも１つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項４】計算機を利用して構造化された種文書に類似する文書を検索する方法であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストのうち指定された構造に属するテキストから特徴となる文字列を抽出するステップと、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として該特徴文字列に基づく該種文書との類似度を算出するステップと、算出された類似度の高い順に従って表示の優先度を決定するステップとを有することを特徴とする構造指定による類似文書の検索方法。

【請求項５】前記種文書は、表示画面上の指示されたテキストであることを特徴とする請求項１又は請求項２記載の構造指定による類似文書の検索方法。

【請求項６】前記検索対象とする構造が指定される代わりに、前記検索対象から除外されるべき構造が指定されることを特徴とする請求項１又は請求項２記載の構造指定による類似文書の検索方法。

【請求項７】前記種文書に属する構造が指定される代わりに、前記種文書に属し検索対象から除外されるべき構造が指定されることを特徴とする請求項３又は請求項４記載の構造指定による類似文書の検索方法。

【請求項８】検索対象として複数の構造が指定されたとき、前記検索対象構造ごとに類似度を算出し、すべての検索対象構造に亘る類似度の累積値を前記優先度決定の際の最終の類似度とすることを特徴とする請求項２又は請求項４記載の構造指定による類似文書の検索方法。

【請求項９】検索対象として複数の構造が指定されたとき、前記検索対象構造ごとに類似度を算出し、対象とする文書について最も高い類似度を前記優先度決定の際の最終の類似度とすることを特徴とする請求項２又は請求項４記載の構造指定による類似文書の検索方法。

【請求項１０】前記特徴文字列を抽出した後に、さらに前記特徴文字列の重要度に応じて採用する特徴文字列を決定することを特徴とする請求項２又は請求項４記載の構造指定による類似文書の検索方法。

【請求項１１】前記特徴文字列を抽出した後に、さらに前記指定された種文書に属する構造の重要度に応じて採用する特徴文字列を決定することを特徴とする請求項４記載の構造指定による類似文書の検索方法。

【請求項１２】前記特徴文字列を抽出した後に、さらに前記検索対象構造の重要度に応じて採用する特徴文字列を決定することを特徴とする請求項２又は請求項４記載の構造指定による類似文書の検索方法。

【請求項１３】検索結果を表示するに際して、表示する文書について前記検索対象構造ごとに抽出された前記特徴文字列を強調表示することを特徴とする請求項２又は請求項４記載の構造指定による類似文書の検索方法。

【請求項１４】検索結果を表示するに際して、表示する文書について前記検索対象構造ごとに前記類似度を表示することを特徴とする請求項２又は請求項４記載の構造指定による類似文書の検索方法。

【請求項１５】検索条件として指定された文書あるいは文章（種文書）に類似する構造化文書を検索する装置であって、類似度計算の検索条件として種文書と該構造化文書に属する少なくとも１つの構造の指定を受ける手段と、類似度計算の後、類似度のより高い対象文書を優先して表示する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項１６】種文書に類似する構造化文書を検索する装置であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストから特徴となる文字列を抽出する手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として、該特徴文字列に基づく該種文書との類似度を算出する手段と、算出された類似度の高い順に従って表示の優先度を決定する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項１７】構造化された種文書に類似する文書を検索する装置であって、種文書と該種文書に属する少なくとも１つの構造の指定を受ける手段と、類似度計算の後、類似度のより高い対象文書を優先して表示する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項１８】構造化された種文書に類似する文書を検索する装置であって、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストのうち指

定された構造に属するテキストから特徴となる文字列を抽出する手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として該特徴文字列に基づく該種文書との類似度を算出する手段と、算出された類似度の高い順に従って表示の優先度を決定する手段とを有することを特徴とする構造指定による類似文書の検索装置。

【請求項１９】計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、検索条件として指定された文書あるいは文章（種文書）に類似する構造化文書を検索するプログラムであり、類似度計算の検索条件として種文書と該構造化文書に属する少なくとも１つの構造の指定を受けるプログラム手段と、類似度計算の後、類似度のより高い対象文書を優先して表示するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【請求項２０】計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、種文書に類似する構造化文書を検索するプログラムであり、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストから特徴となる文字列を抽出するプログラム手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として、該特徴文字列に基づく該種文書との類似度を算出するプログラム手段と、算出された類似度の高い順に従って表示の優先度を決定するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【請求項２１】計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、構造化された種文書に類似する文書を検索するプログラムであり、種文書と該種文書に属する少なくとも１つの構造の指定を受けるプログラム手段と、類似度計算の後、類似度のより高い対象文書を優先して表示するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【請求項２２】計算機読み取り可能な記憶媒体に格納されたプログラムであって、該プログラムは、構造化された種文書に類似する文書を検索するプログラムであり、種文書と検索対象とする構造が指定されたとき、指定された該種文書のテキストのうち指定された構造に属するテキストから特徴となる文字列を抽出するプログラム手段と、抽出された特徴文字列と指定された検索対象構造とが合致する文書を対象として該特徴文字列に基づく該種文書との類似度を算出するプログラム手段と、算出された類似度の高い順に従って表示の優先度を決定するプログラム手段とを有することを特徴とするプログラムを格納する記憶媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、検索条件として指定された文書（種文書）に類似する文書を検索する装置及び方法に係わり、特に構造化文書の構造を対象として検索を行う装置及び方法に関する。

【０００２】

【従来の技術】近年、パーソナルコンピュータやインターネット等の普及に伴い、データベースに蓄積される電子化文書の数が増大しており、膨大な電子化文書の中からユーザが所望する情報を含んだ文書を検索精度よく、高速かつ効率的に検索したいという要求が高まっている。

【０００３】このような要求に対して種々の検索技術が提案されている。例えば特開平１０－２４０７５２号公報によれば、文書を構成する個々の論理的な構造要素が識別できる文書（以下、構造化文書と呼ぶ）を対象として、論理構造に関する条件を検索条件中に付加した検索を行うことにより、精度の高い検索を行うことができる。

【０００４】また特開平１１－１４３９０２号公報は、ユーザが自分の所望する内容の文書あるいは文章（以下、種文書と呼ぶ）を指定し、その文書と類似する文書を検索する類似文書検索技術を開示する。この技術によれば、サンプル文書を示すだけで目的の文書を簡単に検索でき、ユーザが複雑な検索条件式を考えたり入力する手間が省け、効率的な検索ができる。

【０００５】

【発明が解決しようとする課題】上記の特開平１１－１４３９０２号公報の技術によれば、ユーザは使い勝手がよく効率的な検索ができるが、以下に例示するように検索精度の問題を残している。

【０００６】図３は、従来の類似文書検索システムの処理手順を示す図である。検索条件取得プログラムは、検索条件を入力するためのガイダンス画面を表示装置上に表示する。例えば種文書を含む複数の候補文書の文書番号や見出しなどの一覧情報を表示する。検索条件として種文書が指定されると、特徴ｎ－ｇｒａｍ抽出プログラムが起動され、文書ファイルから種文書のテキスト全文を取り出し、テキスト中から特徴文字列を抽出する。次に類似度算出プログラムが起動され、特徴文字列に対応して文書番号とその特徴文字列の出現回数が登録してある出現頻度ファイルを参照し、種文書の特徴文字列に基づいて同じ特徴文字列を使用する関連文書の種文書に対する類似度を算出して候補文書の文書番号、類似度、見出しなどの一覧情報を検索結果として表示する。

【０００７】図３の検索結果によれば、種文書とよく類似する文書は文書４であるにもかかわらず、特徴文字列の出現頻度がより高いために関連の薄い文書１の類似度がより高くなり、優先的に表示されるという問題がある。

【０００８】本発明の目的は、複雑な検索条件の入力を避けるが検索精度のよい類似文書の検索装置及び検索方

法を提供することにある。

【０００９】

【課題を解決するための手段】本発明は、計算機を利用して種文書に類似する構造化文書を検索する方法であって、類似度計算の検索条件として種文書と構造化文書に属する少なくとも１つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有する構造指定による類似文書の検索方法を特徴とする。

【００１０】また本発明は、構造化された種文書に類似する文書を検索する方法であって、種文書とその種文書に属する少なくとも１つの構造の指定を受けるステップと、類似度計算の後、類似度のより高い対象文書を優先して表示するステップとを有する構造指定による類似文書の検索方法を特徴とする。

【００１１】さらに本発明は、上記の機能を備える検索装置を特徴とする。

【００１２】なおここで構造を指定するとは、文書を構成する論理的な構造要素の名称を指定することを意味する。

【００１３】

【発明の実施の形態】以下、本発明の実施形態について図面を用いて説明する。

【００１４】図１は、第１の実施形態の類似文書検索システムの構成図である。本システムを実現する計算機ハードウェアは、表示装置１００、入力装置１０１、中央処理装置（ＣＰＵ）１０２、外部記憶装置１０３、フロッピィディスクドライブ（ＦＤＤ）１０４、主メモリ１０６とこれら装置間を接続するバス１０７から構成される。

【００１５】外部記憶装置１０３は、テキスト１８０、出現頻度ファイル１８１及び構造インデクス１８２を格納する。テキスト１８０は、構造化文書ファイルあるいは構造化されていない文書ファイルの集合を格納する。ここで構造化文書とは、ＳＧＭＬ，ＸＭＬなどの標準形式に準拠した論理構造をもつ文書、あるいは各論理構造ごとに抽出された複数のフラットテキストから構成されるものである。ＦＤＤ１０４を介してフロッピィディスク１０５に格納されている文書が主メモリ１０６を経由してテキスト１８０に登録される。

【００１６】主メモリ１０６に格納されるシステム制御プログラム１１０は、オペレーティングシステム、グラフィカル・ユーザインタフェースを提供するプログラムなどを含む。文書登録制御プログラム１１１は、文書登録用のプログラムの実行を制御する。登録プログラムには、テキスト登録プログラム１２０、出現頻度計数プログラム１４０を含む出現頻度ファイル作成プログラム１２１及び構造インデクス作成プログラム１２２がある。テキスト登録プログラム１２０は、フロッピィディスク１０５上の文書をテキスト１８０に登録するプログラムである。

【００１７】検索制御プログラム１１２は、類似文書の検索に係わるプログラムの実行を制御するプログラムである。検索用のプログラムには、検索条件式解析プログラム１３０、類似文書検索プログラム１３１及び検索結果出力プログラム１３２がある。類似文書検索プログラム１３１には、特徴文字列抽出プログラム１５０、検索対象構造ＩＤ取得プログラム１５１及び類似度算出プログラム１５２が含まれる。これら検索用プログラムの機能については、以下の検索処理手順の説明の中で説明する。検索制御プログラム１１２及び検索用プログラムを記憶媒体に格納し、駆動装置を介して主メモリ１０６に読み込み、ＣＰＵ１０２によって実行することが可能である。

【００１８】主メモリ１０６中に格納される共有ライブラリ１６０として、構造化文書解析プログラム１７０がある。またワークエリア１６１は、テキスト１８０、出現頻度ファイル１８１、構造インデクス１８２から読み込んだデータ等の一時記憶領域や作業用領域として使用される領域である。

【００１９】出現頻度ファイル１８１は、図２の一部に示すように文字列又は単語に対応して、文書番号、その文字列が含まれる論理構造のＩＤ及びその論理構造中の出現回数を格納する。出現頻度ファイル１８１に登録される対象文書は、構造化文書または構造化していない文書である。出現頻度ファイル作成プログラム１２１は、テキスト１８０中の文書を１つずつ読み込み、テキスト中から特徴文字列を抽出し、出現頻度計数プログラム１４０によって各論理構造ごとの特徴文字列の出現回数を計数し、出現頻度ファイル１８１を作成して外部記憶装置１０３に登録する。構造化していない文書については、論理構造の区分が指定されると、その指定に従って特徴文字列と論理構造ＩＤとを対応づける。例えば特開平１１－１４３９０２号公報は出現頻度ファイル作成プログラムの処理手順を開示する。

【００２０】構造インデクス１８２は、図２の一部に示すように論理構造とそのＩＤの対応関係を格納する。構造インデクス作成プログラム１２２は、構造化文書解析プログラム１７０を呼び出し、テキスト１８０から読み込んだ文書テキストの論理構造を解析して各論理構造にＩＤを付与して外部記憶装置１０３に登録する。例えば特開平１０－２４０７５２号公報は、構造インデクス作成プログラムの処理手順を開示する。

【００２１】図２は、第１の実施形態の処理手順を示す図である。第１の実施形態では、種文書が構造化されていない文書、検索対象文書が出現頻度ファイル１８１に登録済みの文書（構造化文書又は構造化されていない文書）とする。検索条件式解析プログラム１３０は、表示装置１００上にガイダンス画面を表示し、検索条件式の入力を受け付ける。

【００２２】ここで検索条件式は、種文書及び少なくとも１つの検索対象構造である。種文書はすでに見出し等が表示された複数の文書候補のうちの１つを選択することが可能であるし、入力装置１０１から直接入力することも可能であるし、ＦＤＤ１０４やＣＤ－ＲＯＭ装置（図には示していない）、ネットワーク（図には示していない）等を介して入力することも可能である。

【００２３】さらに図１２に示すように、表示装置１００に種文書入力用領域１２００、検索対象構造入力用領域１２０１および検索実行ボタン１２０２を備えた画面インタフェースを介して検索条件式が入力されるものとしてもよい。種文書入力用領域１２００には、入力装置１０１より種文書を直接入力することも可能であるし、あるいは検索結果表示画面（図には示していない）上のテキストを種文書入力用領域１２００にコピーすることも可能である。あるいは種文書はこのようなテキストのうちの指示された部分であってもよい。

【００２４】また検索対象構造は、表示されるドロップダウンメニューから少なくとも１つを選択することが可能である。複数の検索対象構造が指定された場合の検索条件では、各構造に対して重みを付与することが可能である。ここで重みは、重み入力用領域（図には示していない）を介して入力されるものでもよいし、システム定義ファイル（図には示していない）で定義されるものとしてもよい。

【００２５】入力装置１０１を介して種文書及び検索対象構造が入力されると、検索条件式解析プログラム１３０は指定された検索条件から種文書のテキストを取得する。なお検索対象とする構造を指定する代わりに、検索対象としない構造（検索対象から除外する構造）を指定してもよい。その場合には、検索条件式解析プログラム１３０は、残りの構造を検索対象構造とする。また検索対象構造を検索条件式の１つとして入力装置１０１を介して入力する代わりにあらかじめシステム定義ファイル（図には示していない）に設定された検索対象構造を用いてもよい。

【００２６】検索対象構造ＩＤ取得プログラム１５１は、構造インデクス１８２を参照して指定された構造に対応する識別子を検索対象構造ＩＤとして取得する。また特徴文字列抽出プログラム１５０は、テキスト１８０から指定された種文書のテキスト全文を取り出し、特徴文字列を抽出し、抽出した特徴文字列の出現回数を計数する。特徴文字列の抽出方法としては、例えば特開平１１－１４３９０２号公報に記載された方法を用いることができる。図２の例では抽出した特徴文字列のうち優先度の高いものを採用している。あるいは文書テキストから単語を切り出し、単語辞書（図には示していない）を参照して登録された単語との一致をチェックしながら単語を抽出してもよい。

【００２７】次に類似度算出プログラム１５２は、出現頻度ファイル１８１を参照して抽出された各特徴文字列又は単語と検索対象構造ＩＤが一致する文書の文書番号とその出現頻度を取得する。次に出現頻度ファイル１８１を参照して取得した各文書の検索対象構造ＩＤについて抽出された特徴文字列又は単語以外の他の文字列又は単語の出現頻度を取得し、各文書ごとに種文書との類似度を算出する。類似度算出方法としては、例えば特開平１１－１４３９０２号公報に記載の数式１を用いることができる。あるいは種文書の各特徴文字列（単語）の正規化された出現ウェイトを要素とする特徴ベクトルと取得した各文書の特徴ベクトルを求め、種文書と他文書の特徴ベクトルの内積によって各文書の類似度を計算してもよい。

【００２８】最後に検索結果出力プログラム１３２は、取得した文書を類似度の高い順に並べ替え、類似度の高い順に従って表示の優先度を決定し、優先度の高い文書から順に文書番号とその類似度を表示装置１００上に表示する。ファイル（図には示していない）を参照して各文書の見出し、概要などの書誌事項を取得して併せて表示してもよい。類似文書との比較のために種文書の文書番号、類似度、見出しなどを併せて表示することも可能である。

【００２９】なお複数の検索対象構造が指定された場合に、各文書の類似度を算出するに際して、各検索対象構造の類似度を全体に亘って累積する累積値を求め、文書をこの類似度の累積値の大きい順に並べ替えてもよい。ここで累積値とは、各検索対象構造ごとの類似度の総和、２乗和の平方根を求めたものなどである。あるいは各文書について複数の検索対象構造の各類似度のうち最も高い類似度を採用し、文書をこの採用した類似度の大きい順に並べ替えてもよい。例えば特許明細書中の［請求項ｎ］のように文書中に同一種類の論理構造が繰り返し出現する場合に、各論理構造ごとに類似度を算出し、その中で最も高い類似度を採用して種文書の類似度と比較すると、同一種類の論理構造の順番には無関係に内容の類似度の高い論理構造同志の比較をすることができる。また類似度の累積値を求めるモードと、最も高い類似度を採用するモードの両方を設け、検索条件の１つとしていずれかのモードを選択できるようにしてもよいし、あらかじめシステム定義ファイル（図には示していない）に選択するモードを設定できるようにしてもよい。

【００３０】図４は、第２の実施形態の類似文書検索プログラム１３１ａの構成を示す図である。第２の実施形態では、特徴文字列抽出プログラム１５０ａに種文書構造解析プログラム４００が加わっている。種文書構造解析プログラム４００は、共有ライブラリ１６０に格納されている構造化文書解析プログラム１７０を呼び出す構成をとる。また類似度算出プログラム１５２ａに対応構造判定プログラム４０１が加わっている。

【００３１】図５は、第２の実施形態の処理手順を示す図である。第２の実施形態では種文書が構造化文書、検索対象文書が出現頻度ファイル１８１に登録済みの文書（構造化文書又は構造化されていない文書）とする。第２の実施形態の検索条件は種文書及び種文書に属する少なくとも１つの構造である。入力装置１０１を介して種文書及び構造が入力されると、検索条件式解析プログラム１３０は指定された検索条件から種文書のテキストを取得する。検索条件で指定された種文書の論理構造と検索対象文書の論理構造が一致するものとする。なお検索対象とする構造を指定する代わりに、検索対象としない構造（検索対象から除外する構造）を指定してもよい。その場合には、検索条件式解析プログラム１３０は、種文書に属する残りの構造を検索対象構造とする。第１の実施形態と同様に検索対象構造をあらかじめシステム定義ファイルに設定しておいてもよい。

【００３２】次に種文書構造解析プログラム４００は、テキスト１８０から種文書のテキストを取り出し、種文書の構造を解析して指定された構造に関する本文テキストのみを抽出する。種文書のテキストに指定された構造が含まれていないときにはエラーとする。文書の構造解析の方法としては、例えば特開平１０－２４０７５２号公報に文書構造解析プログラムの処理手順として記載されている。次に検索対象構造ＩＤ取得プログラム１５１は、構造インデクス１８２を参照して指定された構造に対応する検索対象構造ＩＤを取得する。また特徴文字列抽出プログラム１５０ａは、抽出されたテキストの特徴文字列（単語）を抽出し、抽出した特徴文字列の出現回数を計数する。なお種文書について、すでに各構造の特徴文字列が抽出され、その構造ごとの出現回数が計数されており、出現頻度ファイル１８１のように登録されているのであれば、そのファイルを参照して指定された構造、特徴文字列と出現回数を抽出するだけでよい。この場合には種文書構造解析プログラム４００及び特徴文字列抽出プログラム１５０ａの処理をスキップできる。

【００３３】次に類似度算出プログラム１５２ａは、第１の実施形態と同様に出現頻度ファイル１８１を参照して抽出された各特徴文字列（単語）と検索対象構造ＩＤが一致する文書の文書番号を取得し、各文書ごとに種文書との類似度を算出する。この際に対応構造判定プログラム４０１は、種文書の構造と、出現頻度ファイル１８１から取得された文書の構造ＩＤとの対応をとり、特徴文字列を検索対象構造ごとのグループに分け、検索対象構造ごとの類似度を算出する。複数の検索対象構造が指定された場合に、各文書の最終的な類似度を算出する方法は第１の実施形態と同様である。最後に検索結果出力プログラム１３２は、取得した文書を算出した類似度の高い順に並べ替えてその文書番号、類似度、見出し等を表示装置１００上に表示する。

【００３４】なお上記の第２の実施形態の説明では、指定された種文書の構造と検索対象構造とが一致するものとしたが、両者が別の論理構造であってもよい。すなわち種文書の構造が指定され、これとは別の検索対象構造が指定された場合、種文書構造解析プログラム４００及び特徴文字列抽出プログラム１５０ａは、指定された種文書の構造に注目して特徴文字列を抽出し、類似度算出プログラム１５２ａは指定された検索対象構造のＩＤに注目して検索対象文書を検索する。また対応構造判定プログラム４０１は、指定された種文書の構造と指定された検索対象構造が同一グループとみなして対応づけをする。例えば薬の効能書の［副作用］を特徴文字列を抽出するときの対象構造とし、［効能］を検索対象文書の類似度を計算するときの検索対象構造とすることにより、種文書に記載の薬のもつ副作用を抑える薬について記載された文書を探し出すことが可能となる。

【００３５】図６は、第２の実施形態の問題点を説明する図である。この例では種文書に関する構造として［効能］［副作用］［使用上の注意］が指定され、これらの構造が検索対象構造とみなされ、検索が実行されている。その結果”服用”“自動車”“運転”など薬の効能書にとって重要度が小さいか無意味な特徴文字列が抽出され、これらの特徴文字列を含む特徴文字列に基づく類似度算出の結果として、文書２、文書３などあまり重要でない文書の類似度が無視できない程の値を示し、検索結果として文書２、文書３などが挙がったことを示している。

【００３６】図７は、第３の実施形態の類似文書検索プログラム１３１ｂの構成を示す図である。第３の実施形態では、特徴文字列抽出プログラム１５０ｂにさらに構造重みプログラム６００が加わっている。

【００３７】図８は、第３の実施形態の処理手順を示す図である。第３の実施形態は、特徴文字列抽出プログラム１５０ａが抽出した特徴文字列に対して構造重みプログラム６００を適用する以外は第２の実施形態の処理と同じである。構造重みプログラム６００は、論理構造ごとに重要度が設定してあるシステム定義ファイルを参照して、各論理構造ごとにその重要度に応じて抽出した特徴文字列の中から検索用として採用する特徴文字列の数を決定する。例えば「重要」の構造は抽出されたすべての特徴文字列を採用し、「普通」の構造は抽出された特徴文字列の重要度に従って一部の特徴文字列のみを採用する。あるいは各論理構造ごとに採用する特徴文字列の数をシステム定義ファイルに設定し、抽出された特徴文字列からその重要度が上位の所定数の特徴文字列を採用してもよい。また所定の文字種からそれぞれ所定数を採用するようにしてもよい。特徴文字列の重要度を算出する方法としては、例えば特開平１１－１４３９０２号公報は数式２として特徴文字列の重要度の算出式を挙げている。なお各論理構造の重要度や特徴文字列の採用個数をシステム定義ファイルに設定する代わりに、検索条件

式の一部として入力装置１０１を介して指定してもよい。なお論理構造の重要度あるいは特徴文字列の重要度により採用する特徴文字列を決定する方式は、上記の第１の実施形態にも適用可能である。

【００３８】以上のようにして特徴文字列を絞り込んだ上で類似度算出プログラム１５２ａ及び対応構造判定プログラム４０１を適用すると、検索結果から重要度の少ない文書を排除することができる。また第１、第２の実施形態に比べて特徴文字列の数が削減されることになるので、出現頻度ファイル１８１を検索する際の検索時間を短縮できる。

【００３９】図９は、第４の実施形態の検索結果表示プログラム１３２ａの構成を示す図である。第４の実施形態では、検索結果表示プログラム１３２ａに構造別表示方法取得プログラム７００が加わっている。

【００４０】図１０は、第４の実施形態の処理手順を示す図である。第４の実施形態は、検索結果表示プログラム１３２ａの処理を除いては第１～第３の実施形態の処理と同じである。構造別表示方法取得プログラム７００は、類似度算出プログラム１５２ａの処理結果として挙げられた文書について検索対象構造別の類似度を表示する。また検索対象構造ごとに抽出された特徴文字列を強調表示する。

【００４１】図１１は、構造別表示方法取得プログラム７００の処理手順を示すＰＡＤ図である。構造別表示方法取得プログラム７００は、特徴文字列抽出プログラム１５０ａにより抽出された各構造ごとの特徴文字列をそれぞれワークエリア１６１に格納する（ステップ７０１）。次に類似度算出プログラム１５２ａにより算出された各構造ごとの類似度をワークエリア１６１に格納する（ステップ７０２）。次に検索された各文書の指定されたすべての構造について以下の処理を繰り返す（ステップ７０３）。まずワークエリア１６１に格納された当該構造の類似度を取得し表示する（ステップ７０４）。次にワークエリア１６１に格納された当該構造の特徴文字列を取得し、強調表示する（ステップ７０５）。なおこの実施形態では各論理構造ごとに類似度と特徴文字列の強調表示とを行うものとしたが、いずれか一方のみを行ってもよい。また検索結果の表示条件をシステム定義ファイル上に設定してもよいし、検索条件式の一部として指定してもよい。

【００４２】なお種文書及び検索結果として挙げられた類似文書について、各々文書番号に対応して見出し、概要などを表示し、これらのテキストに含まれ、採用された特徴文字列を強調表示してもよい。このように表示すると、類似文書に含まれる特徴文字列を種文書に含まれる特徴文字列と比較することができる。

【００４３】なお上記実施形態で使用した出現頻度ファイル１８１の代わりに特開平１１－１４３９０２号公報のｎ－ｇｒａｍインデクスを用いてもよい。すなわち特開平１１－１４３９０２号公報と特開平１０－２４０７５２号公報の構造化された文字列インデックスを組み合わせると、出現頻度ファイル１８１に代わるファイルを構成可能である。同一文字列についての１つ以上の出現位置はその文字列の出現回数をも示している。

【００４４】なお上記第１～第４の実施形態では種文書として１つの文書が指定されるものとしたが、複数の種文書を指定できるものとしてもよい。ここで特徴文字列としては、それぞれの種文書から抽出された特徴文字列をすべて用いるものとしてもよいし、それぞれの種文書に共通して含まれる特徴文字列を用いるものとしてもよい。

【００４５】

【発明の効果】以上述べたように本発明によれば、類似文書検索の検索条件として論理構造の指定を付加するので、類似文書検索の利点を最大限に生かしながら検索精度を高めることができる。なお複数の検索対象構造が指定される場合に、あらかじめ設定された論理構造の重要度または特徴文字列の重要度に応じて関連の薄い特徴文字列を排除でき、さらに検索精度を高めることができる。また種文書は、構造化文書と構造化していない文書のいずれも可能であり、ユーザが種文書の選択に注意を払う必要がない。

【図面の簡単な説明】

【図１】実施形態の類似文書検索システムの構成図である。

【図２】第１の実施形態の処理手順を示す図である。

【図３】従来の類似文書検索システムの処理手順を示す図である。

【図４】第２の実施形態の類似文書検索プログラム１３１ａの構成を示す図である。

【図５】第２の実施形態の処理手順を示す図である。

【図６】第２の実施形態の問題点を説明する図である。

【図７】第３の実施形態の類似文書検索プログラム１３１ｂの構成を示す図である。

【図８】第３の実施形態の処理手順を示す図である。

【図９】第４の実施形態の検索結果出力プログラム１３２ａの構成を示す図である。

【図１０】第４の実施形態の処理手順を示す図である。

【図１１】第４の実施形態の構造別表示方法取得プログラムの処理手順を示す図である。

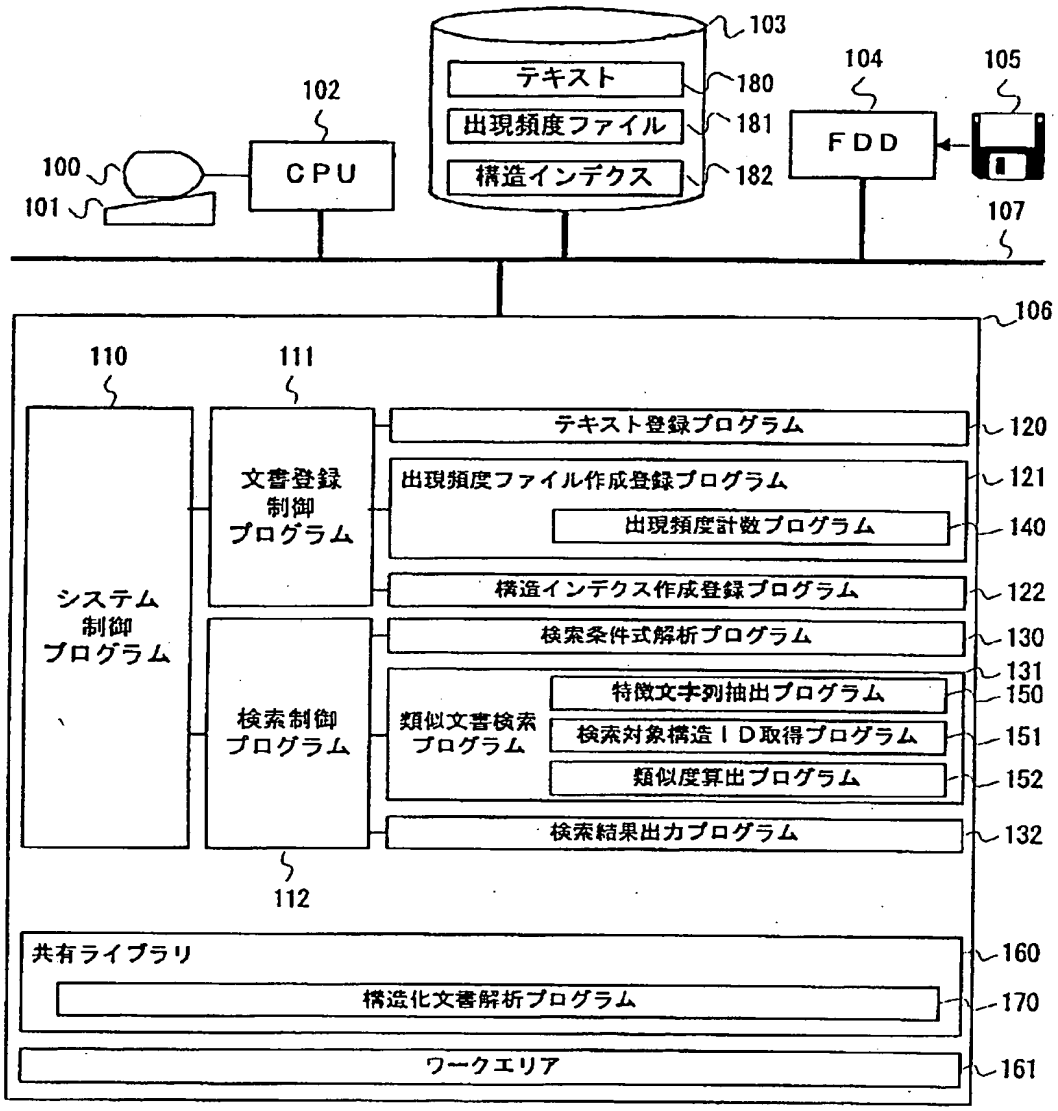
【図１２】検索条件入力画面の例を示す図である。

【符号の説明】

１３１：類似文書検索プログラム、１３２：検索結果出力プログラム、１５０：特徴文字列抽出プログラム、１５１：検索対象構造ＩＤ取得プログラム、１５２：類似度算出プログラム、１８０：テキスト、１８１：出現頻度ファイル、１８２：構造インデクス、４００：種文書構造解析プログラム

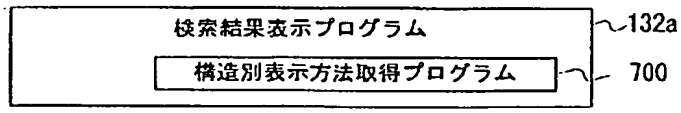
【図１】

図１


103
テキスト
180
出現頻度ファイル
181
構造インデクス
182
102
CPU
100
101
104
FDD
105
107
106
110
111
システム
制御
プログラム
文書登録
制御
プログラム
テキスト登録プログラム
120
出現頻度ファイル作成登録プログラム
121
出現頻度計数プログラム
140
構造インデクス作成登録プログラム
122
検索制御
プログラム
検索条件式解析プログラム
130
131
類似文書検索
プログラム
特徴文字列抽出プログラム
150
検索対象構造ＩＤ取得プログラム
151
類似度算出プログラム
152
検索結果出力プログラム
132
112
共有ライブラリ
160
構造化文書解析プログラム
170
ワークエリア
161

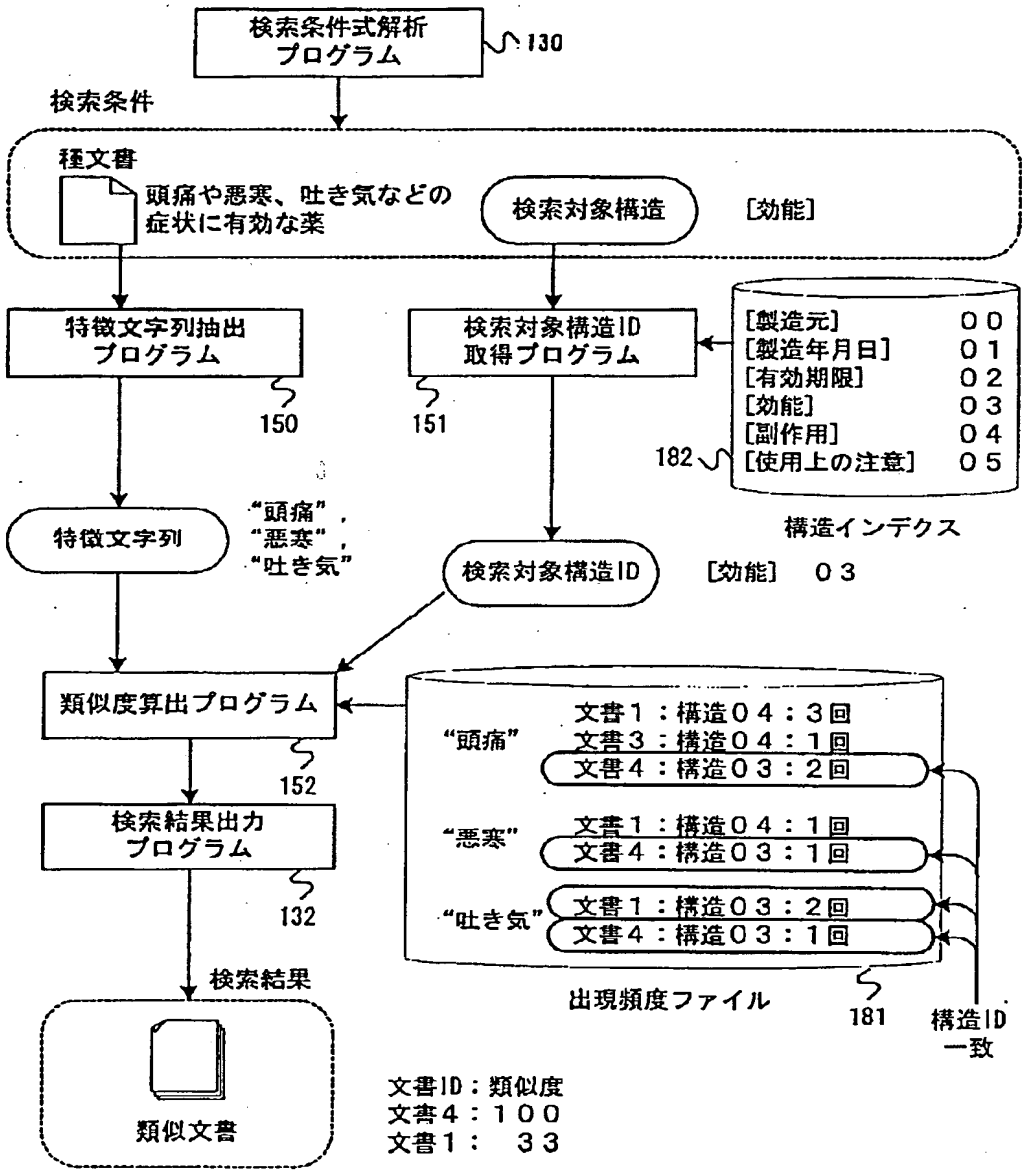
【図９】

図９


検索結果表示プログラム
132a
構造別表示方法取得プログラム
700

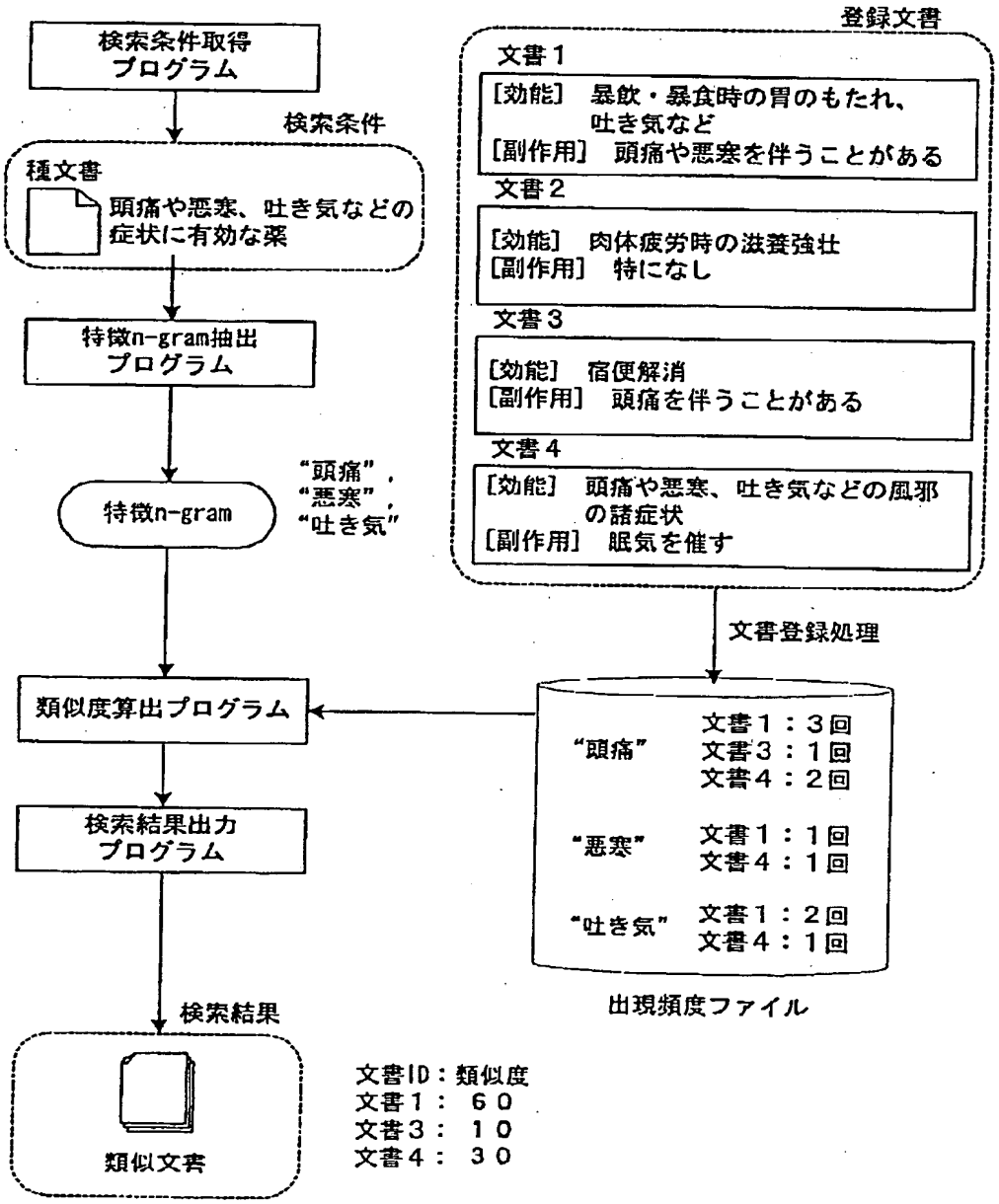
【図２】

図２


検索条件式解析
プログラム
130
検索条件
種文書
頭痛や悪寒、吐き気などの
症状に有効な薬
検索対象構造
[効能]
特徴文字列抽出
プログラム
150
検索対象構造ID
取得プログラム
151
[製造元]　００
[製造年月日]　０１
[有効期限]　０２
[効能]　０３
[副作用]　０４
[使用上の注意]　０５
182
構造インデクス
特徴文字列
“頭痛”,
“悪寒”,
“吐き気”
検索対象構造ID
[効能]　０３
類似度算出プログラム
152
検索結果出力
プログラム
132
検索結果
類似文書
“頭痛”
文書１：構造０４：３回
文書３：構造０４：１回
文書４：構造０３：２回
“悪寒”
文書１：構造０４：１回
文書４：構造０３：１回
“吐き気”
文書１：構造０３：２回
文書４：構造０３：１回
出現頻度ファイル
181
構造ID
一致
文書ID：類似度
文書４：１００
文書１：　３３

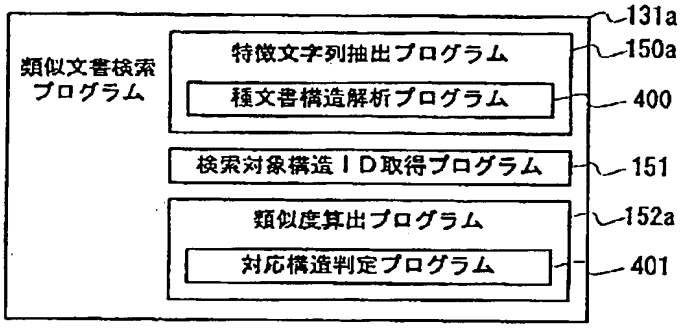
【図３】

図３


検索条件取得
プログラム
検索条件
種文書
頭痛や悪寒、吐き気などの
症状に有効な薬
特徴n-gram抽出
プログラム
特徴n-gram
“頭痛”，
“悪寒”，
“吐き気”
類似度算出プログラム
検索結果出力
プログラム
検索結果
類似文書
文書ID：類似度
文書１：　６０
文書３：　１０
文書４：　３０
登録文書
文書１
[効能]　暴飲・暴食時の胃のもたれ、
吐き気など
[副作用]　頭痛や悪寒を伴うことがある
文書２
[効能]　肉体疲労時の滋養強壮
[副作用]　特になし
文書３
[効能]　宿便解消
[副作用]　頭痛を伴うことがある
文書４
[効能]　頭痛や悪寒、吐き気などの風邪
の諸症状
[副作用]　眠気を催す
文書登録処理
“頭痛”　文書１：３回
文書３：１回
文書４：２回
“悪寒”　文書１：１回
文書４：１回
“吐き気”　文書１：２回
文書４：１回
出現頻度ファイル

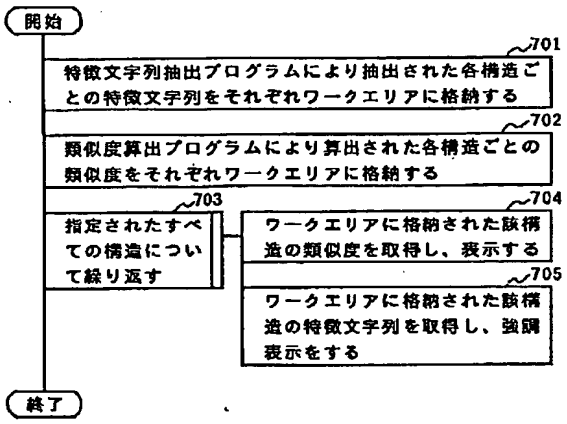
【図４】

図４



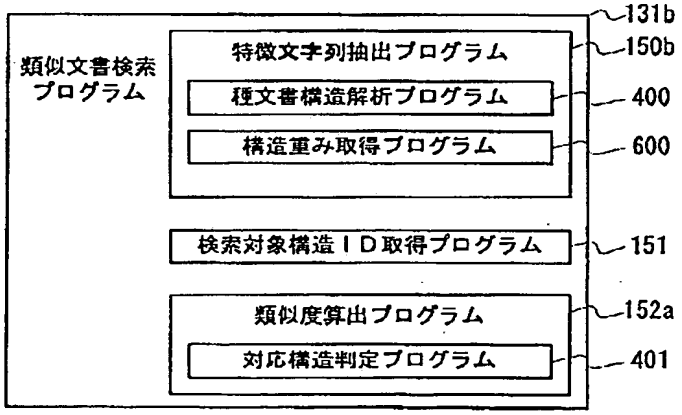
【図１１】

図１１



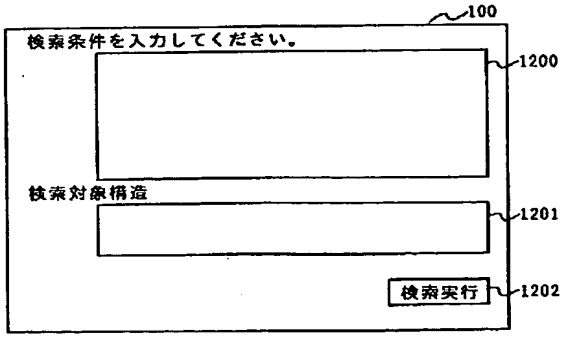
【図７】

図７



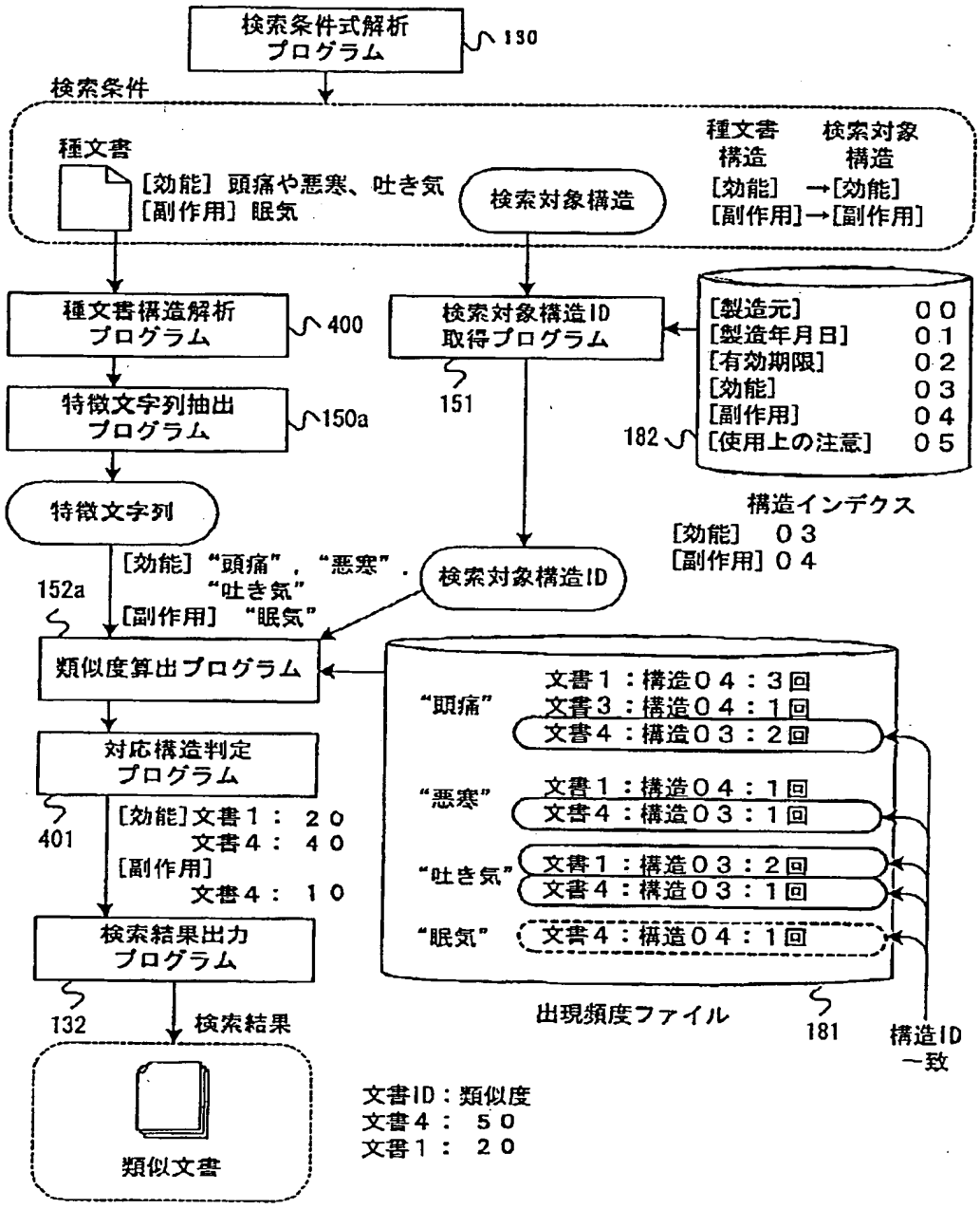
【図１２】

図１２



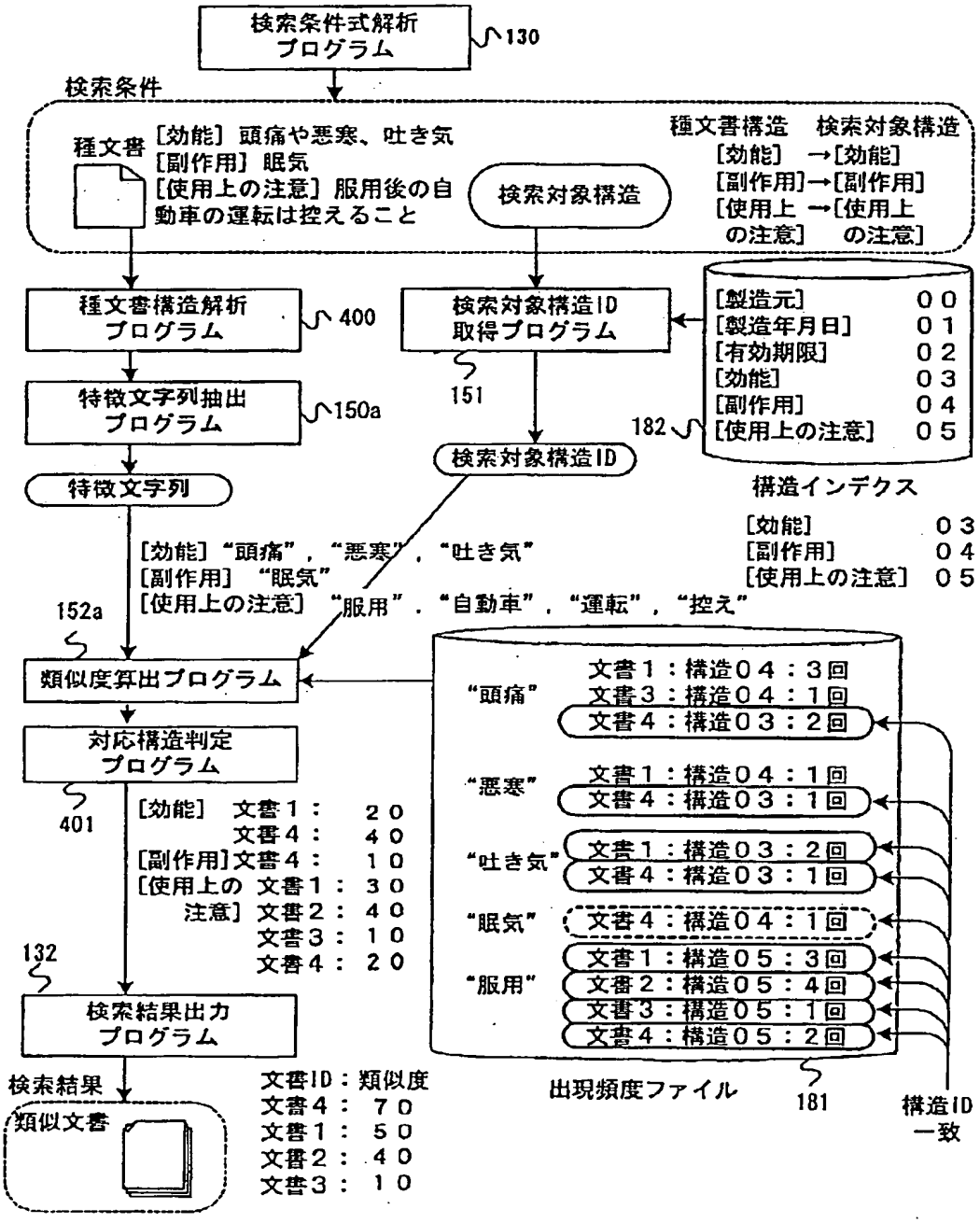
【図５】

図５


検索条件式解析プログラム
130
検索条件
種文書
[効能] 頭痛や悪寒、吐き気
[副作用] 眠気
検索対象構造
種文書構造　検索対象構造
[効能] →[効能]
[副作用]→[副作用]
種文書構造解析プログラム
400
検索対象構造ID取得プログラム
151
特徴文字列抽出プログラム
150a
特徴文字列
[効能]“頭痛”，“悪寒”，“吐き気”
[副作用] “眠気”
152a
検索対象構造ID
類似度算出プログラム
対応構造判定プログラム
401
[効能]文書１： ２０
文書４： ４０
[副作用]
文書４： １０
検索結果出力プログラム
132
検索結果
類似文書
文書ID：類似度
文書４： ５０
文書１： ２０
[製造元] ００
[製造年月日] ０１
[有効期限] ０２
[効能] ０３
[副作用] ０４
[使用上の注意] ０５
182
構造インデクス
[効能] ０３
[副作用] ０４
“頭痛” 文書１：構造０４：３回
文書３：構造０４：１回
文書４：構造０３：２回
“悪寒” 文書１：構造０４：１回
文書４：構造０３：１回
“吐き気” 文書１：構造０３：２回
文書４：構造０３：１回
“眠気” 文書４：構造０４：１回
出現頻度ファイル
181
構造ID一致

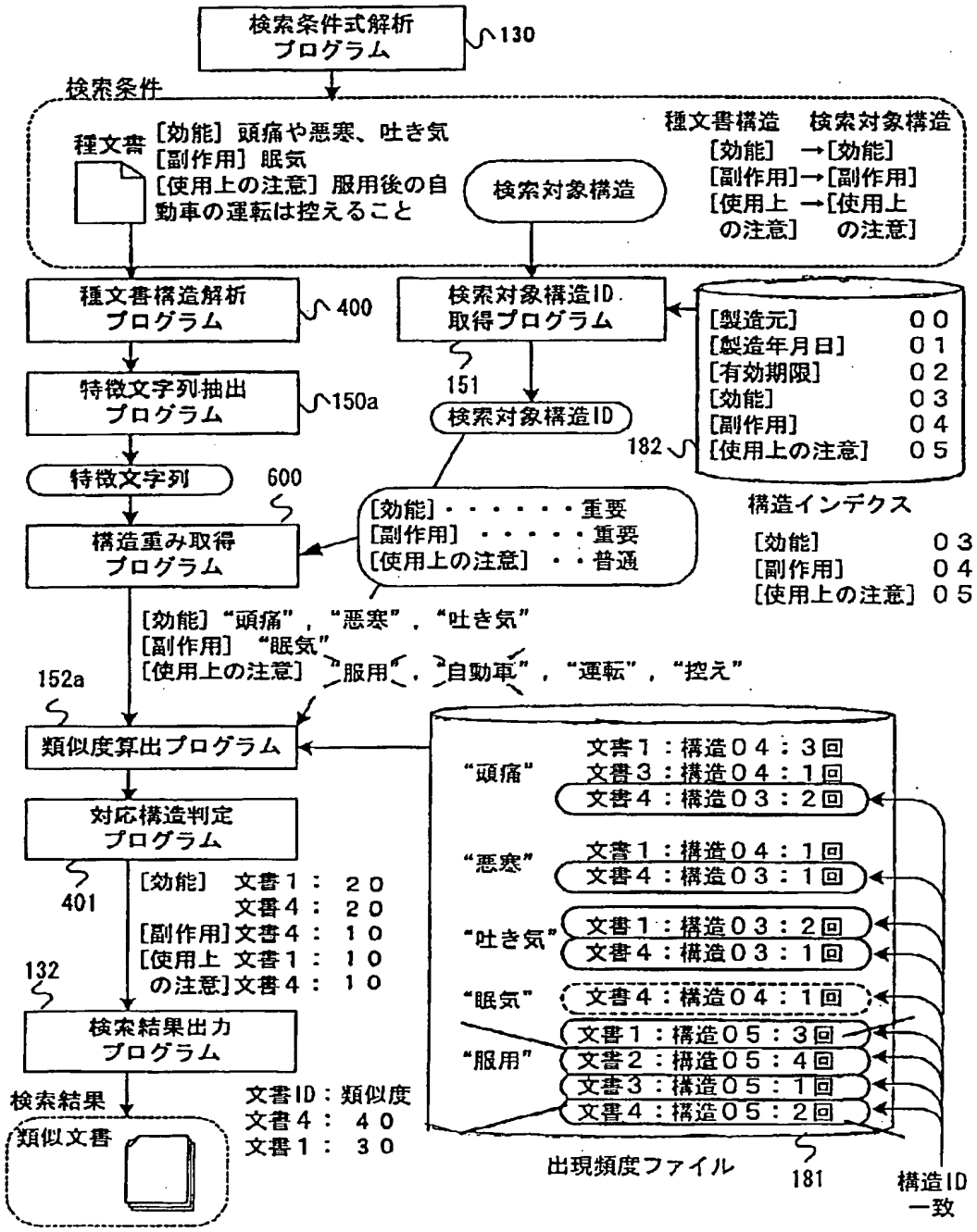
【図６】

図６


検索条件式解析プログラム
130
検索条件
種文書
[効能] 頭痛や悪寒、吐き気
[副作用] 眠気
[使用上の注意] 服用後の自動車の運転は控えること
検索対象構造
種文書構造　検索対象構造
[効能] →[効能]
[副作用]→[副作用]
[使用上の注意] →[使用上の注意]
種文書構造解析プログラム
400
検索対象構造ID取得プログラム
151
[製造元] 00
[製造年月日] 01
[有効期限] 02
[効能] 03
[副作用] 04
[使用上の注意] 05
182
構造インデクス
特徴文字列抽出プログラム
150a
検索対象構造ID
特徴文字列
[効能] "頭痛"，"悪寒"，"吐き気"
[副作用] "眠気"
[使用上の注意] "服用"，"自動車"，"運転"，"控え"
[効能] 03
[副作用] 04
[使用上の注意] 05
152a
類似度算出プログラム
対応構造判定プログラム
401
[効能] 文書１：２０
文書４：４０
[副作用] 文書４：１０
[使用上の注意] 文書１：３０
文書２：４０
文書３：１０
文書４：２０
132
検索結果出力プログラム
検索結果
類似文書
文書ID：類似度
文書４：７０
文書１：５０
文書２：４０
文書３：１０
"頭痛" 文書１：構造０４：３回 文書３：構造０４：１回 文書４：構造０３：２回
"悪寒" 文書１：構造０４：１回 文書４：構造０３：１回
"吐き気" 文書１：構造０３：２回 文書４：構造０３：１回
"眠気" 文書４：構造０４：１回
"服用" 文書１：構造０５：３回 文書２：構造０５：４回 文書３：構造０５：１回 文書４：構造０５：２回
出現頻度ファイル
181
構造ID一致

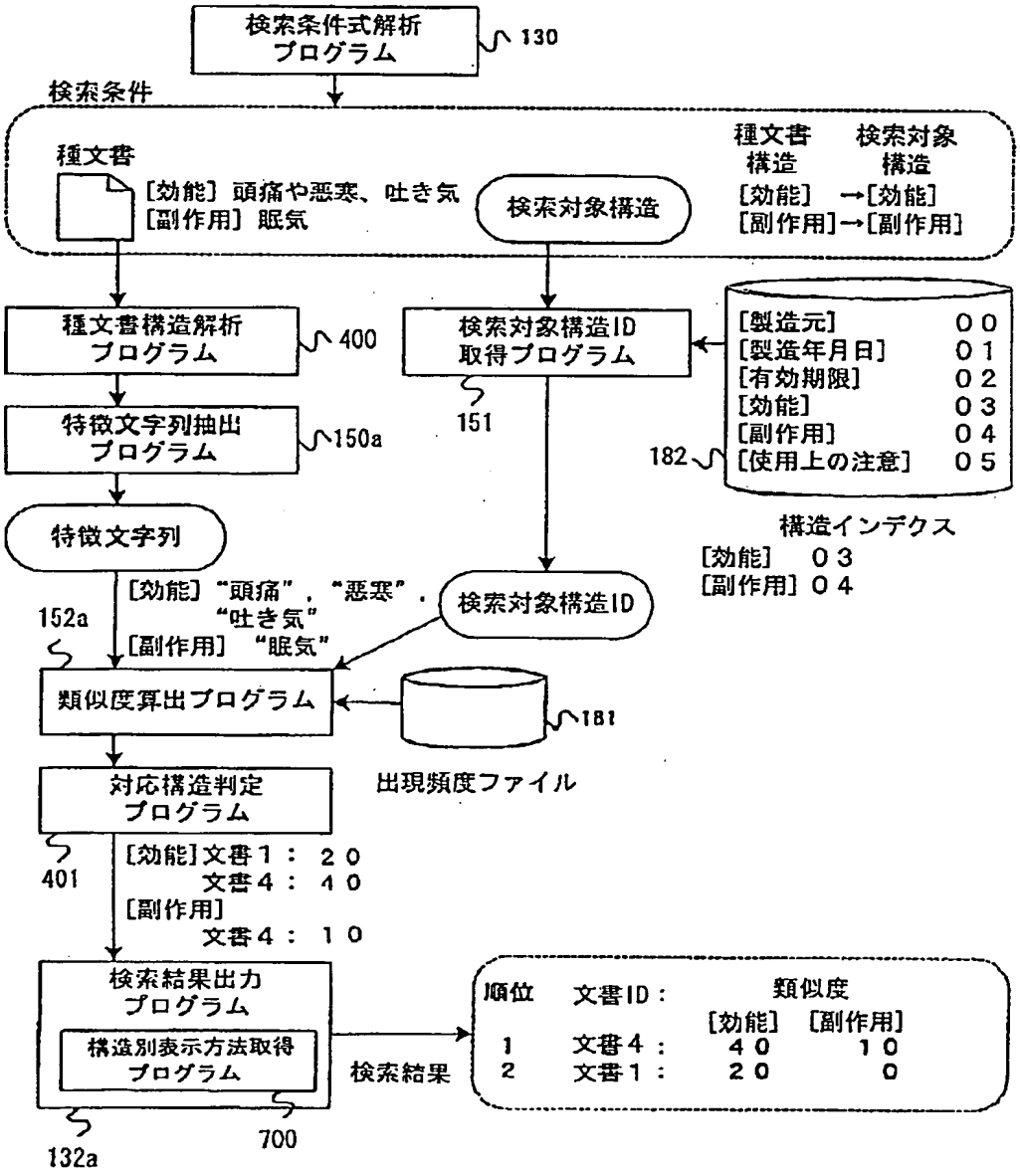
【図８】

図８


検索条件式解析プログラム
130
検索条件
種文書
[効能] 頭痛や悪寒、吐き気
[副作用] 眠気
[使用上の注意] 服用後の自動車の運転は控えること
検索対象構造
種文書構造　検索対象構造
[効能] →[効能]
[副作用]→[副作用]
[使用上の注意] →[使用上の注意]
種文書構造解析プログラム
400
検索対象構造ID取得プログラム
151
特徴文字列抽出プログラム
150a
検索対象構造ID
特徴文字列
600
構造重み取得プログラム
[効能]・・・・・・重要
[副作用]・・・・・・重要
[使用上の注意]・・普通
[製造元] ００
[製造年月日] ０１
[有効期限] ０２
[効能] ０３
[副作用] ０４
[使用上の注意] ０５
182
構造インデクス
[効能] ０３
[副作用] ０４
[使用上の注意] ０５
[効能]“頭痛”，“悪寒”，“吐き気”
[副作用] “眠気”
[使用上の注意] “服用”，“自動車”，“運転”，“控え”
152a
類似度算出プログラム
対応構造判定プログラム
401
[効能] 文書１： ２０
文書４： ２０
[副作用] 文書４： １０
[使用上の注意] 文書１： １０
文書４： １０
132
検索結果出力プログラム
検索結果
類似文書
文書ID：類似度
文書４： ４０
文書１： ３０
“頭痛”
文書１：構造０４：３回
文書３：構造０４：１回
文書４：構造０３：２回
“悪寒”
文書１：構造０４：１回
文書４：構造０３：１回
“吐き気”
文書１：構造０３：２回
文書４：構造０３：１回
“眠気”
文書４：構造０４：１回
“服用”
文書１：構造０５：３回
文書２：構造０５：４回
文書３：構造０５：１回
文書４：構造０５：２回
出現頻度ファイル
181
構造ID一致

【図１０】

図１０



フロントページの続き

(72)発明者　菅谷　奈津子
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(72)発明者　稲場　靖彦
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(72)発明者　山口　明彦
神奈川県川崎市幸区鹿島田890番地　株式会社日立製作所システム開発本部内
(72)発明者　後地　陽介
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア事業部内

Ｆターム(参考) 5B009 QA09 VA02
5B075 ND03 NK06 NK39 PP13 PQ02
PQ22 PQ36 PQ46 PQ75 PR06
QM08